

Beberapa Konsep dalam Adaptasi Tes

Wahyu Widhiarso | Fakultas Psikologi UGM

FAIRNESS

A group of hands holding up large red letters spelling 'FAIRNESS'. The letters are thick and three-dimensional. The hands are of various skin tones and are positioned at the bottom of each letter, holding them up from below. The background is white, and the entire scene is framed by a thin black border.

Awal Cerita ...



Universal design is an approach to test design that seeks to maximize accessibility for all intended examinees.

- Aksesibilitas menunjukkan bahwa semua peserta tes harus memiliki kesempatan yang sama untuk menunjukkan diri mereka terkait konstruk yang sedang diukur" (p .49).
- Desain universal didefinisikan sebagai "pendekatan untuk menguji desain yang berupaya memaksimalkan aksesibilitas untuk semua peserta ujian yang dimaksud" (p. 50).

Universal Design

- Equitable Use
- Flexibility in Use
- Simple and Intuitive Use
- Perceptible Information
- Tolerance for Error
- Low Physical Effort
- Size and Space for Approach and Use

Table 9.1 Applying Universal Test Design Principles

Universal Design Principles							Universal Test Design Principles	Observable Characteristics
Equitable Use	Flexibility in Use	Simple and Intuitive Use	Perceptible Information	Tolerance for Error	Low Physical Effort	Size and Space for Approach and Use		
X	X						1. Inclusive assessment population Opportunity for participation for all members of the target population regardless of physical characteristics, culture, linguistic background, or cognitive abilities.	This information is difficult to “observe,” but most teacher-developed assessments are consistent with this principle.
X				X			2. Precisely defined constructs Performance should not be affected by construct-irrelevant variance, processes that are extraneous to the intended construct (AERA et al., 1999).	Points awarded for knowledge or performance, not construct-irrelevant tasks (e.g., speed, handwriting, perhaps spelling and grammar). Wording for math problems should be simple and clear.
X	X						3. Accessible, nonbiased items Items are biased if groups of equal ability have different probabilities of answering correctly. Items also should be free of culturally offensive content.	Words, phrases, and concepts are commonly used across cultures and languages. No pop culture references (e.g., TV, music). No stereotypes or offensive terms.

Universal Design Principles							Universal Test Design Principles	Observable Characteristics
Equitable Use	Flexibility in Use	Simple and Intuitive Use	Perceptible Information	Tolerance for Error	Low Physical Effort	Size and Space for Approach and Use		
X	X		X			X	4. Amenable to accommodations The way in which a test is presented can easily be changed to remove unintended disadvantages for English language learners or for those with disabilities.	Horizontal text. No construct-irrelevant graphs or pictures. Graphics are simple and clear. Keys and legends at top or right of item. No time limits. Subsections of tests are independent of each other.
X		X	X				5. Simple, clear, and intuitive instructions and procedures “Assessment instructions and procedures need to be easy to understand, regardless of a student’s experience, knowledge (or) language skills . . .” (Thompson et al., 2002, p. 14)	Consistent instructions (e.g., circling correct answer). Directions allow students to work independently without questions. Practice or sample items are provided. Numbered items.

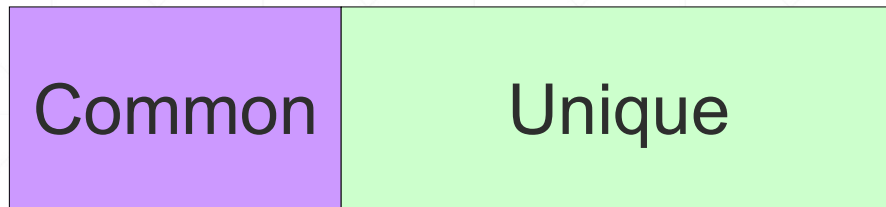
- Universal design avoids, where possible, item characteristics and formats, or test characteristics (for example, inappropriate test speededness), that may bias scores for individuals or subgroups due to **construct-irrelevant characteristics** that are specific to these test takers (p.58).

Construct underrepresentation: The extent to which a test fails to capture important aspects of the construct domain that the test is intended to measure, resulting in test scores that do not fully represent that construct (p. 217).

Construct-irrelevant variance: “Variance in test-taker scores attributable to extraneous factors that distort the meaning of the scores, thereby decreasing the validity of the proposed interpretation (p. 217).

	Orang#1	Orang#2	Orang#3
1. Saya suka berteman	[Ya]	[Tidak]	[Ya]
2. Saya suka menyapa	[Ya]	[Tidak]	[Ya]
3. Saya suka bermain	[Tidak]	[Tidak]	[Ya]
	(+)	(-)	(+)

Total Variance



Common

Unique

**Total
Variance**

**Common
Variance**

**Unique
Variance**

Sistematik

Random

Construct
Irrelevant
Variance

Common
Variance

Konstruk
Ukur

**Total
Variance**

X1

X2

X3

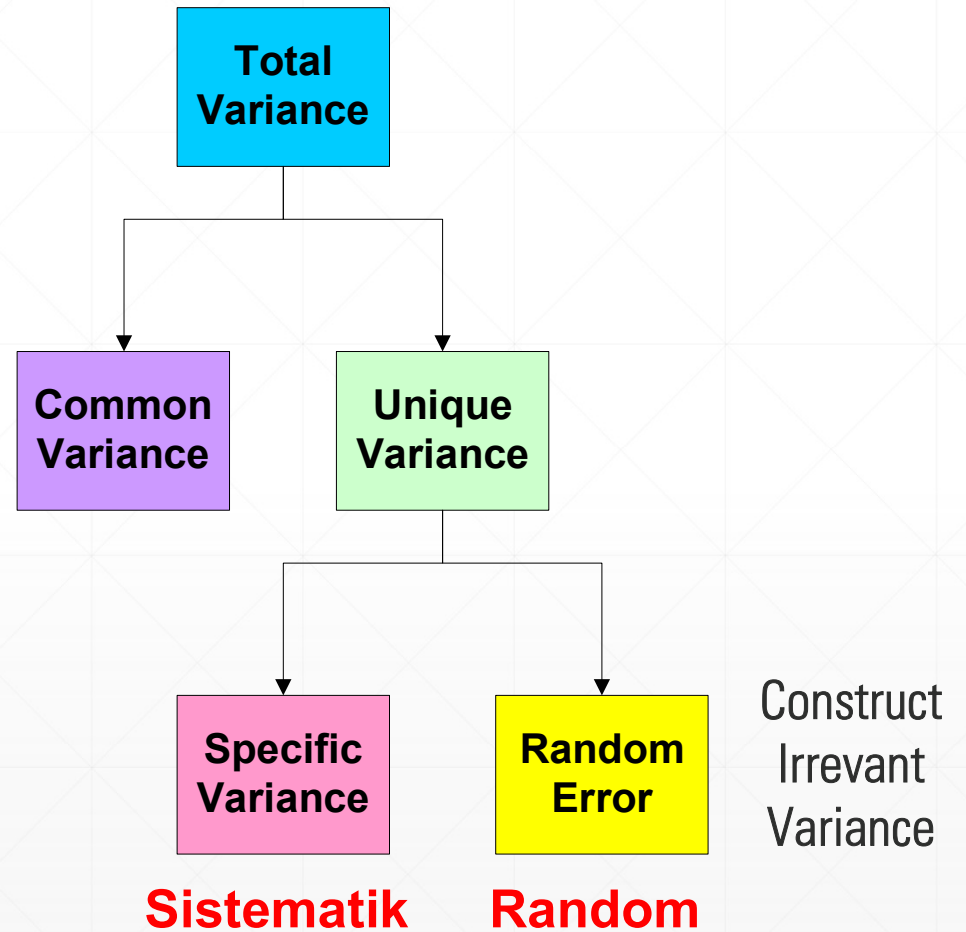
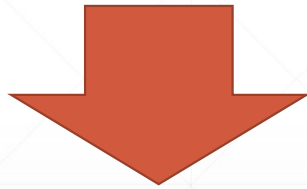
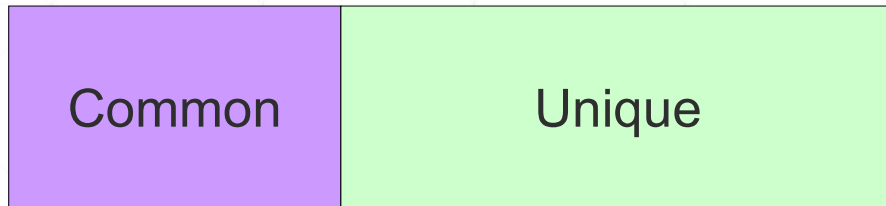
X4

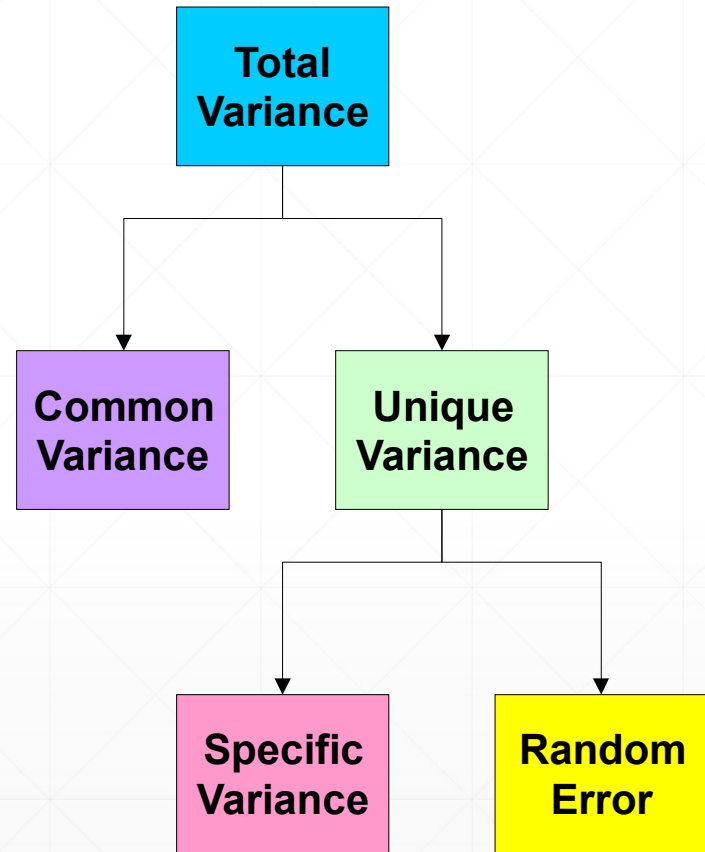
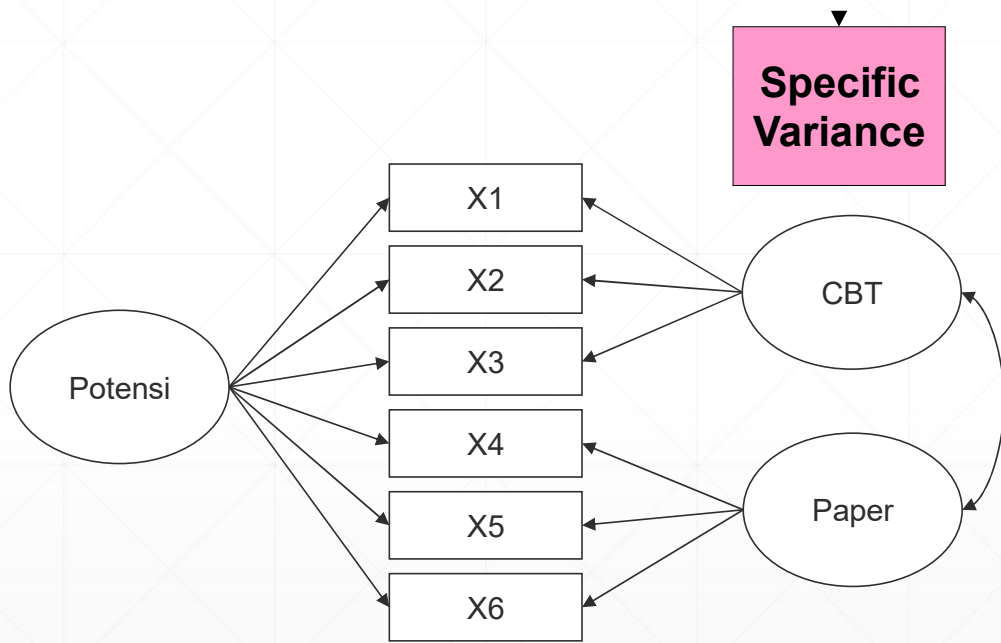
X5

X6

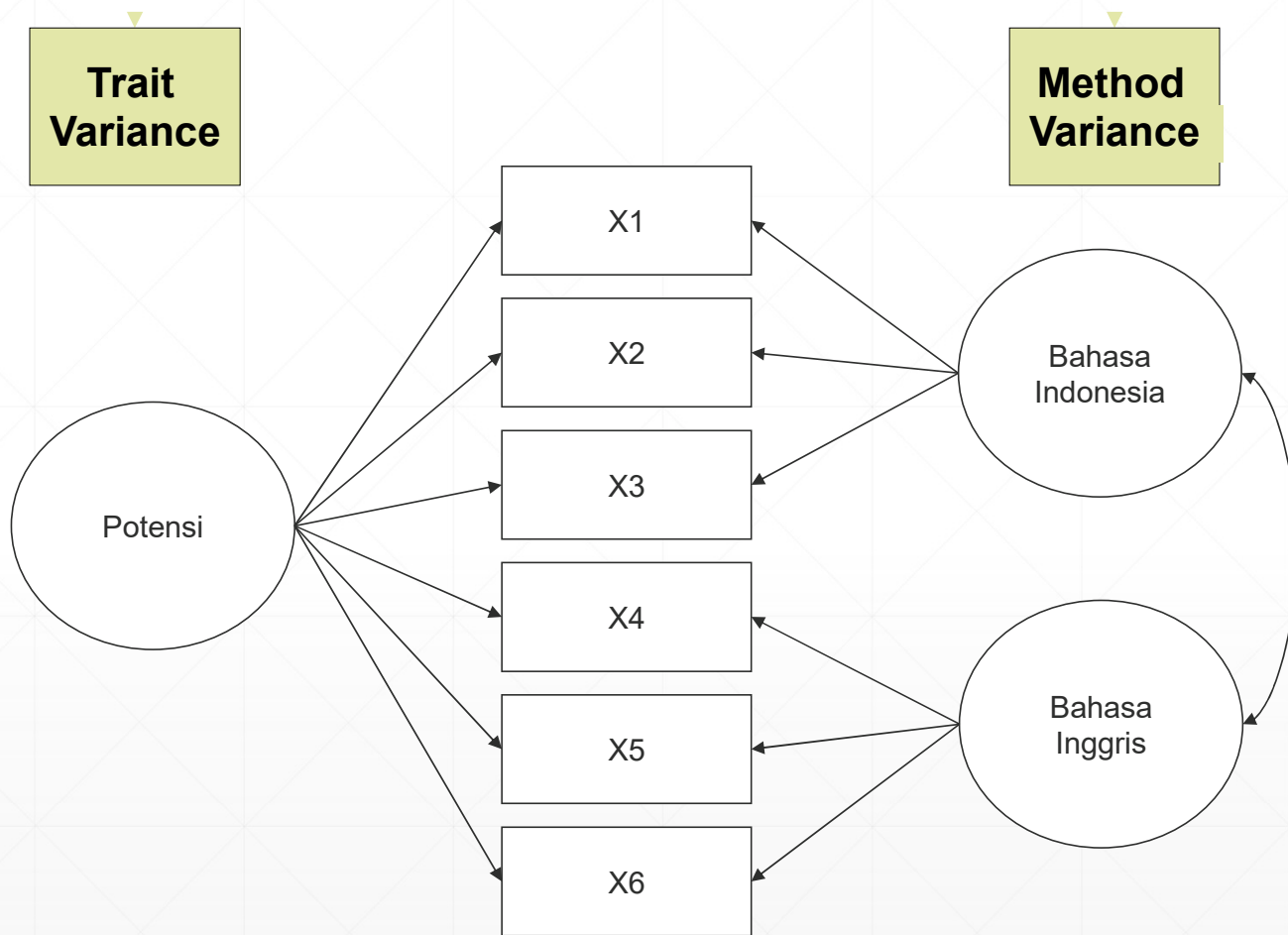
Common
Variance

Total Variance





Sistematik **Random**
Construct Irrelevant Variance



adaptasi

Mengapa ada Adaptasi?

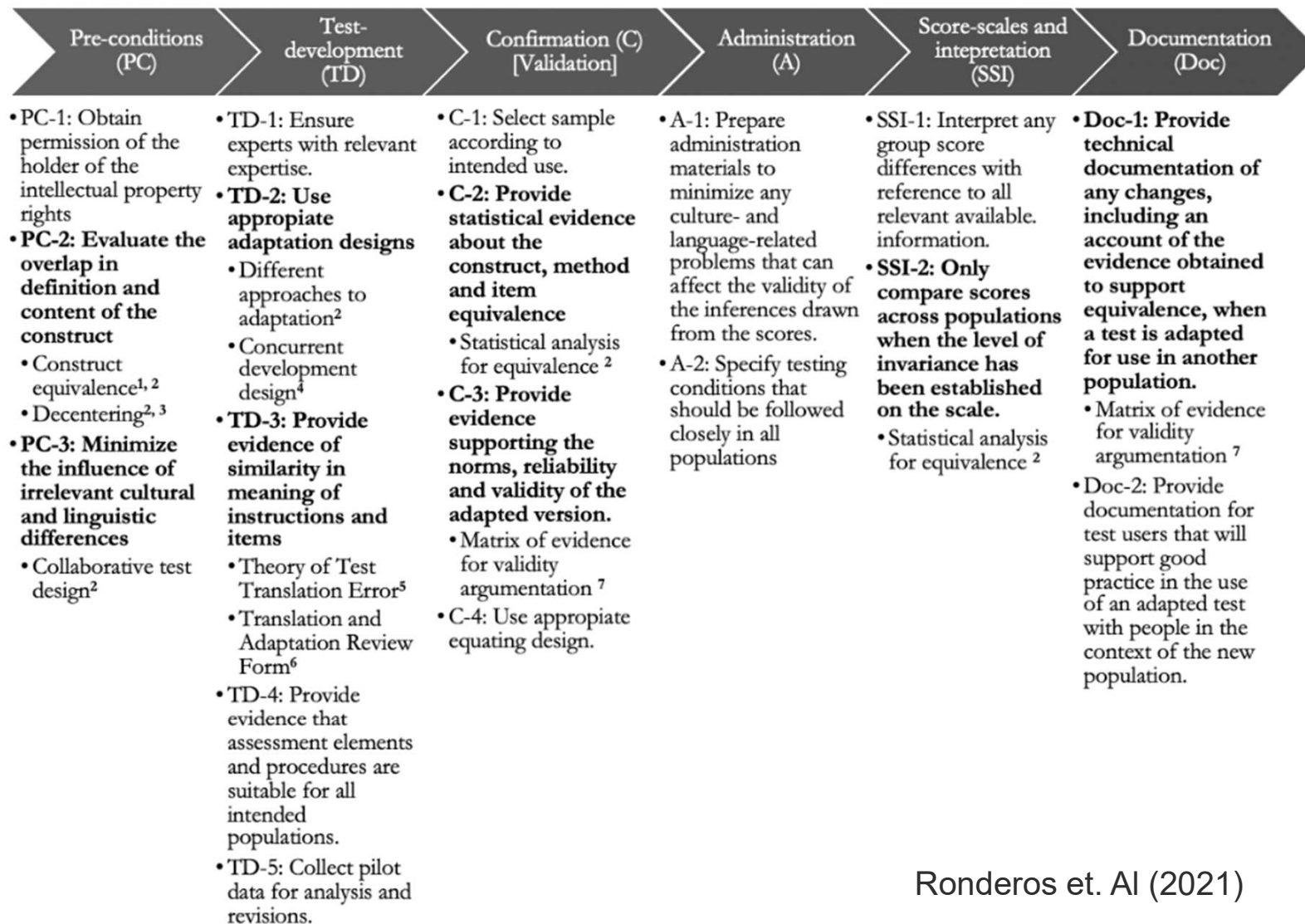


Even when tests are developed to **maximize fairness** through the use of **universal design** and other practices to increase access, there will still be situations where the test is not appropriate for all test takers in the intended population. Therefore, **some test adaptations may be needed** for those individuals whose characteristics would otherwise impede their access to the examination (APA, AERA, NCME, 2014)

Karena desain universal belum cukup ...

Adaptasi

- Adaptasi
 - Any change in test content, format, or administration conditions that is made to increase a tests accessibility for individuals who otherwise would face construct-irrelevant barriers on the original test. An adaptation may or may not change the meaning of the construct being measured or alter score interpretations.
- Adaptasi memiliki dua tipe
 - Modifikasi
 - Akomodasi



Ronderos et. Al (2021)

Akomodasi

Adjustments that do not alter the assessed construct that are applied to test presentation, environment, content, format, or administration conditions for particular test takers, and that are embedded within assessments or applied after the assessment is designed.

Tests or assessments with such accommodations, and their scores, are said to be accommodated. Accommodated scores should be sufficiently comparable to unaccommodated scores that they can be aggregated together (p. 215)

Request test accommodations for your exam

Get started



Understanding test accommodations

We provide appropriate arrangements to individuals who demonstrate a documented need. Test accommodations are individualized, so we consider them on a case-by-case basis in partnership with your testing program. Test accommodations may also be known as “special arrangements” or “reasonable adjustments.”

Pearson VUE test accommodations may include:

- Extra testing time
- A separate testing room
- Breaks

Modifikasi

- Changes made in the content, format, and/or administration procedure of a test to increase the accessibility of the test for test takers who are unable to take the original test under standard testing conditions.
- In contrast to test accommodations, test modifications change the construct being measured by the test to some extent and hence change score interpretations.

Contoh Modifikasi

- Standar: mengeja 20 kata → modifikasi: dikurangi menjadi 10 kata
- Standar: pertanyaan tentang pengalaman masa lalu → modifikasi: pertanyaan tentang masa sekarang
- Standar: tidak boleh pakai kalkulator → modifikasi: boleh pakai kalkulator

ACCOMMODATION

**CHANGES HOW A
STUDENT LEARNS**

MODIFICATION

**CHANGES WHAT A
STUDENT LEARNS**

EXAMPLES OF ACCOMMODATIONS

- Using a larger font or print size
- Taking frequent breaks
- Having extra time to complete assignments or tests
- Presenting information in a different way or format
- Using sensory tools & fidgets
- Changing how assignments or tests are completed
- Using visual aids & supports
- Allowing a student to sit where & how they learn best

EXAMPLES OF MODIFICATIONS

- Assigning different projects or assignments
- Using different test questions
- Teaching different material than that taught to peers
- Excusing the student from certain projects or tests
- Using different educational standards or rubrics for grading & assessment

Adaptasi atau Translasi?

- Adaptasi mengacu pada proses yang memuat bahasa, budaya seperti kesetaraan konstruk dan familiaritas dengan format butir (Berman et al., 2020; Ercikan & Por, 2020; van de Vijver & Poortinga, 2016)

Adaptasi atau Membuat Baru?

- But when cross-cultural comparisons are not of interest, it may be easier to actually produce a new test that meets the cultural parameters in the second language group, than to adapt an already existing test which may have a number of shortcomings (Hambleton & Patsula, 2004)

Beberapa Poin

Desain Penelitian

- Bilingual group design,
- Matched monolingual group design
- monolingual group design.

Kesetaraan (Ekuivalensi) Konstruk

Kesetaraan Konstruk (*construct equivalence*)

- The extent to which a construct measured by one test is **essentially the same** as the construct measured by another test. The degree to which a construct measured by a test in one cultural or **linguistic** group is **comparable** to the construct measured by the same test in a different cultural or linguistic group.
- Kesetaraan konstruk secara makna dan kebahasaan
 - Pressure = Stress = Tekanan? | Tension = Arousal = Tegangan | Happiness = kebahagiaan
- Ekuivalen? Tidak → *decentering* atau berhenti
- Hambleton (2005) merekomendasikan proses ini menekankan pada *judgmental approaches*

Perspektif Ketat : Enforced Etic

Instruments for measuring an enforced etic are constructed or adapted for the new culture based on the assumption that the measured construct exists in that culture, although this is yet unknown, but is to be verified (Hedrih, 2020)

- Enforced Etic adalah konstruk psikologis yang belum ditemukan apakah merupakan konstruk yang, tetapi universalitas budayanya sedang diselidiki.
- Instrumen untuk mengukur Enforced Etic dikembangkan atau diadaptasi untuk budaya baru berdasarkan asumsi bahwa konstruk yang diukur eksis pada budaya itu meskipun hal ini belum diketahui, tetapi harus diverifikasi.

Instruksi. Pilihlah tingkat persetujuan Anda dengan dengan pernyataan berikut.

Wajah saya cakep

Wajah saya rupawan

Wajah saya tampan

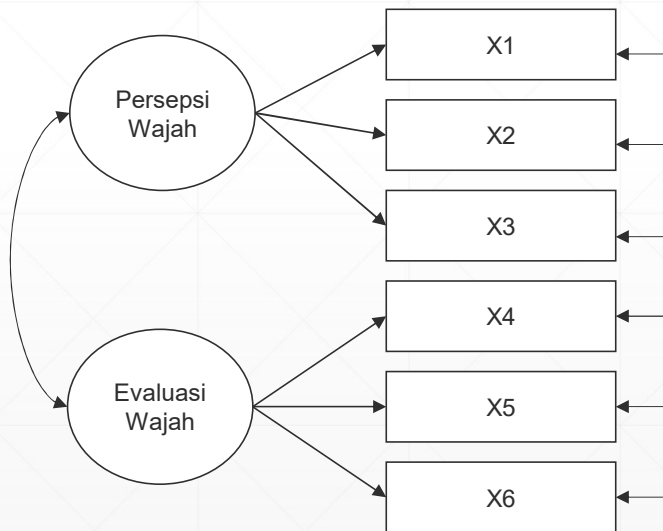
Wajah saya menawan

Wajah saya memesona

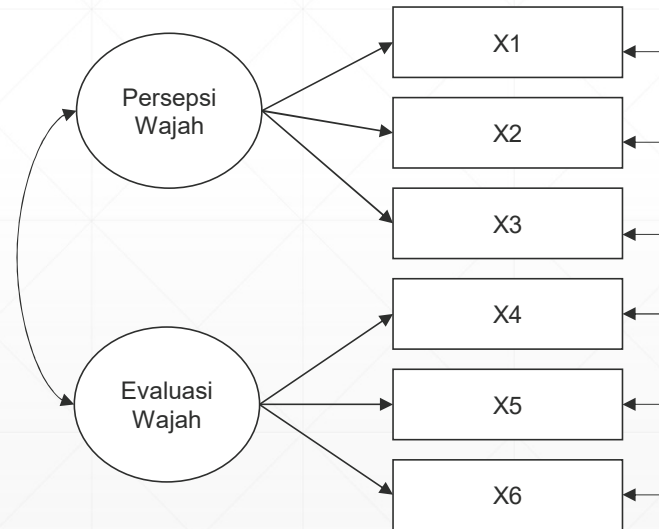
Wajah saya menarik

Perspektif Ketat

Budaya A – Model Fit



Budaya B – Model Fit



Construct Unequivalence

Perspektif Ketat

Construct equivalence is defined as the equivalence of meaning of the construct in terms of its theoretical definition, the way it is operationalized, and the way it is developed for the cultures in which the assessment will be administered. In addition to demonstrating **similar psychometric properties**, the evidence for whether a construct (e.g., reading proficiency) is conceptualized the same way across language and sociocultural groups also needs to be grounded in empirical research based on the considered language and sociocultural groups.

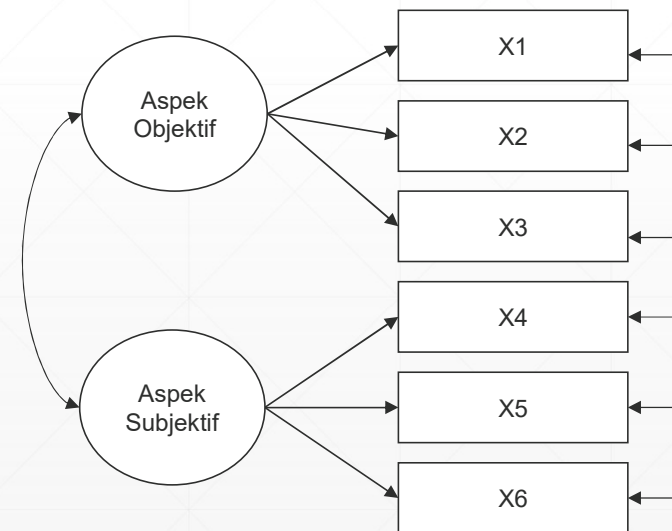
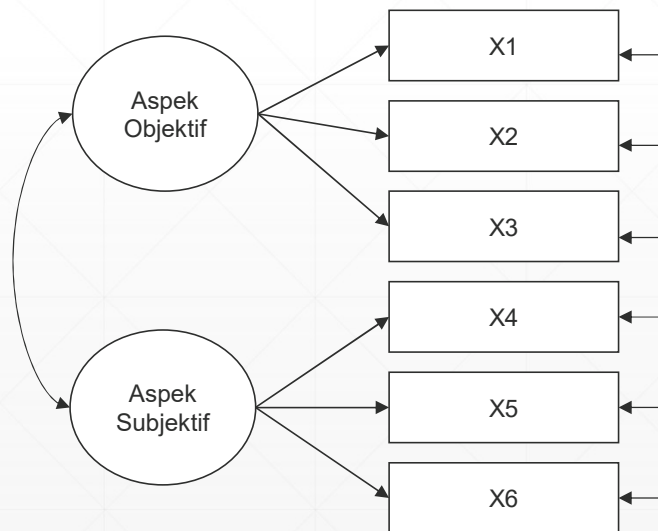
Contoh

The European Spanish adaptation of the PAM-13 **has similar psychometric properties** to the original scale, thus warranting its future application in the clinical and research context in Spain

The confirmatory factor analysis' outputs suggest that the translated version of the MIQ has a good performance in two-dimensional models and display **similar psychometric properties** regarding other population and languages (Black & Reynolds, 2016). Therefore, the fact that no further changes were needed suggests that the Spanish version of the MIQ can be used to assess moral identity in the Spanish population.

Level Ekuivalensi Konstruk

- Konseptual = Kesetaran konstruk
- Structural = Kesamaan struktur (jumlah dimensi)
- Metric = Kesamaan bobot faktor
- Scalar = Kesamaan rerata (intercept)



Kok bisa strukturnya bisa berubah?

- Budaya tertentu melihat orang cakep ya pasti menawan (unidimensi), tapi pada budaya lain belum tentu (multidimensi)

Kok bisa metriknya beda?

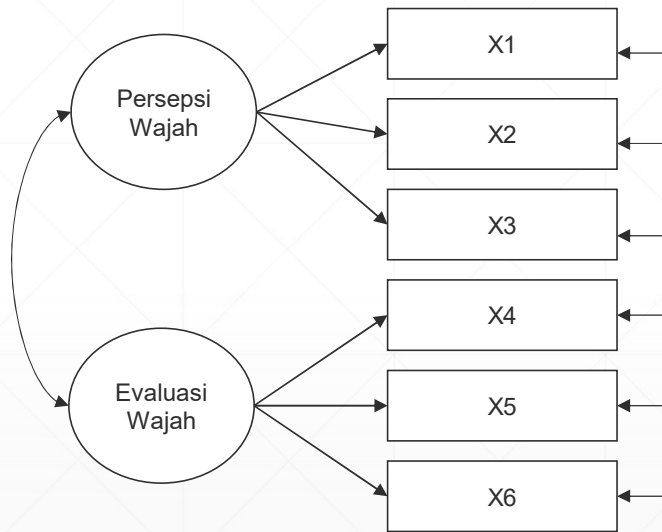
- Budaya A memiliki variasi kecakepan yang lebih tinggi (bobot faktor tinggi), sementara budaya lain kecakepan tidak seberapa bervariasi (bobot faktor rendah)

Kok bisa scalarnya beda?

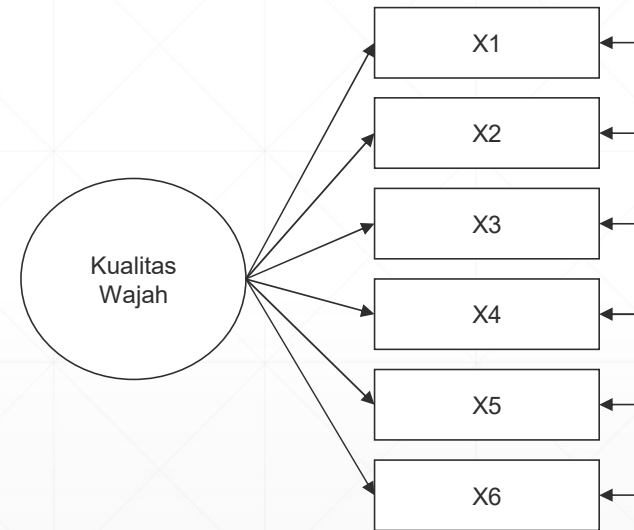
- Budaya tertentu orangnya belum ngisi kuesioner, cakepnya sudah level 2,5 (range 1-5), sementara budaya lain cakepnya baru level 1,2 (skala 1-5)

Pandangan Longgar

Budaya A – Model Fit



Budaya B - Fit



Construct Equivalence

.... does not make the prior assumptions that constructs will be the same across cultures and, consequently, implies a need to establish whether the concept exists and is interpreted similarly. (Herdman, Fox-Rushby, Badia, 1998) – Universalist Perspective

Landasan Pandangan

- Konvergensi, homogenitas atau konsistensi butir merepresentasikan suatu konstruk laten
- Jika bukti-bukti pada versi asli terkait dengan properti psikometris tersebut ditemukan pada versi adaptasi, dapat disimpulkan bahwa yang diukur mirip atau setara dengan versi asli

Table 1
Standardized factor loadings and residuals of items on the GSES-12

Items	Three-factor	
	loading	residual
<i>Initiative</i>		
(1) If something looks too complicated I will not even bother to try it	0.74	0.45
(2) I avoid trying to learn new things when they look too difficult	0.50	0.75
(3) When trying to learn something new, I soon give up if I am not initially successful	0.60	0.63
<i>Effort</i>		
(1) When I make plans, I am certain I can make them work	0.57	0.68
(2) If I can't do a job the first time, I keep trying until I can	0.49	0.76
(3) When I have something unpleasant to do, I stick to it until I finish it	0.53	0.72
(4) When I decide to do something, I go right to work on it	0.56	0.68
(5) Failure just makes me try harder	0.41	0.83
<i>Persistence</i>		
(1) When I set important goals for myself, I rarely achieve them	0.37	0.86
(2) I do not seem capable of dealing with most problems that come up in my life	0.61	0.63
(3) When unexpected problems occur, I don't handle them very well	0.62	0.62
(4) I feel insecure about my ability to do things	0.62	0.62

Ekivalensi Konstruk

Level Ekuivalensi

- Domain-domain di versi asli relevan dengan budaya versi adaptasi, penekanan setiap domain juga sama
- Domain sama, akan tetapi penekanan berbeda → pembobotan
- Satu atau lebih domain versi asli tidak relevan dengan budaya → menggantinya dengan domain yang relevan dengan budaya versi adaptasi
- Domain versi asli tidak relevan dengan budaya partisipan versi adaptasi – Tidak boleh diadaptasi

(Herdman, Fox-Rushby, Badia, 1998)

Ekivalensi Butir

Level Ekivalensi Butir

- Butir soal dipakai tanpa modifikasi
- Butir soal memerlukan modifikasi
- Butir soal memerlukan penggantian
- Butir soal tidak bisa dipakai (masalah tabu dsb.)

Contoh Study #1

Menghasilkan sama-sama 3 faktor namun struktur di dalamnya berbeda

- This study, therefore, investigated the initial validity of the TNT's factor structure and items among Hong Kong youth to determine its cross-cultural validity. Responses to the TNT were gathered from three samples of Hong Kong youth (ages 11 to 18) and subjected to principal components and confirmatory factor analyses. **Results revealed the three-factor TNT solution (i.e., Physical and Psychological Violence; Empathy; Satyagraha) reported in the United States was supported, along with two additional but non-interpretable dimensions** unique to Hong Kong. In general, the findings suggest preliminary evidence for the TNT's cross-cultural validity with Hong Kong adolescents. Implications for research and psycho-educational programs to teach Hong Kong youth conflict resolution strategies are offered.

Contoh Study #2

APA PsycNet Direct

Registered Users Sign In Now

Search APA PsycNet

APA PsycTests: Citation and Description

Conscientiousness Inventory--Modified Version

Widhiarso, W., Steyer, R., & Ravand, H. (2019). *Conscientiousness Inventory--Modified Version (MCI)* [Database record]. APA PsycTests.

<https://doi.org/10.1037/t73197-000>

Description

The Conscientiousness Inventory--Modified Version (MCI; Widhiarso, Steyer, & Ravand, 2019) was developed to measure individuals' level of consciousness using construct-irrelevant item attributes (CIAs) in an employee selection setting. Statements from the original 12-item Conscientiousness Inventory of the NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1989) written with certain restrictions so that the stems embodied seven proposed CIAs. The resulting 64-item MCI was evaluated in a sample of undergraduate students. Factor analysis, reliability, and validity results were reported for these items. (APA PsycTests Database Record (c) 2019 APA, all rights reserved)

Table 1
Item location parameter for each item attribute.

Item attributes	Honest group		Faking group		Wald statistic	N items
	M	SD	M	SD		
Verifiability						
More verifiable	1.34	0.21	1.95	0.97	36.582 (5)*	5/3
Less verifiable	1.10	0.33	0.76	0.78	98.239 (5)*	5/4
Source of inform.						
1st hand	1.11	0.30	0.71	0.22	50.208 (5)*	5/2
2nd hand	1.23	0.14	1.11	0.20	3.582 (5)	5/0
Time frame						
Past	1.22	0.30	1.25	0.28	9.610 (5)	5/1
Future	0.86	0.11	0.13	0.84	144.016 (5)*	5/4
Discreteness						
Non-discrete	1.08	0.12	0.75	0.27	26.828 (4)*	4/2
Discrete	1.35	0.18	1.45	0.71	29.387 (4)*	4/3
Definiteness						
Definite	1.15	0.14	0.91	0.23	19.292 (4)*	4/2
Indefinite	1.19	0.26	0.63	0.20	29.539 (4)*	4/4
Internality						
External	1.31	0.30	1.16	0.46	14.763 (4)*	4/1
Internal	1.10	0.43	0.00	0.36	246.325 (4)*	4/4
Context specificity						
Specific	1.17	0.17	0.17	0.29	151.592 (5)*	5/5
Non-specific	1.20	0.12	1.54	0.46	17.501 (5)*	5/2

Note. N items represents number of items/number of items affected by DIF.

* $p < .001$.

Contoh Study #3

No.	Item content
1	I feel that I'm a person of worth, at least on an equal plane with others.
2	I feel that I have a number of good qualities.
3	All in all, I am inclined to feel that I am a failure.
4	I am able to do things as well as most other people.
5	I feel I do not have much to be proud of.
6	I take a positive attitude toward myself.
7	On the whole, I am satisfied with myself.
8	I wish I could have more respect for myself.
9	I certainly feel useless at times.
10	At times I think I am no good at all.

Satu butir soal tidak dapat dipakai

- Factor loading displayed all items had loading of >0.30 except for item 8 which had negative loading. This indicates moderate correlation between all the items and their respective factors except for item 8 which measures construct opposite to its intended ([Tavakol and Wetzel, 2020](#)). Countries like Bangladesh, the Democratic Republic of the Congo, Indonesia, Malaysia, and Tanzania had negative loadings on item 8 and relatively low loadings were seen in countries like Bolivia, Botswana, Japan, Mexico, Romania, South Korea and Zimbabwe ([Schmitt and Allik, 2005](#)), thus, indicating dubiety in **item 8 original statement**.
- This differential understanding of **item 8 has been presumed to be cultural effect as this finding was commonly seen in collectivistic cultures** ([Keetie Roelen and Emily Taylor, 2020](#)) or among people who have more positive views of themselves ([Schmitt and Allik, 2005](#)). However, some have refuted the cultural effect ([Schmitt and Allik, 2005](#)).

Translasi

- Myth 1. The preferable strategy is always to adapt an existing test rather than develop a new test for a second language group.
- Myth 2. Anyone who knows the two languages can produce an acceptable translation of a test.
- Myth 3. A well-translated test guarantees that the test scores will be valid in a second language or culture for cross-language comparative purposes.
- Myth 4. Constructs are universal, and therefore all tests can be translated into other languages and cultures.
- (Hambleton & Patsula, 2004)

Contoh Hasil Reviu Translasi

- **Changes in difficulty of words or sentences:** The committee felt the translation was accurate, but some words or sentences became easier or more difficult after translation.
- **Changes in content:** The committee felt that the meaning of 8 items was changed in the translation, thus turning it into a different item.
- **Changes in format:** In 5 cases, changes in the format of the item were identified as the probable cause of DIF. For example, a sentence became much longer
- **Differences in cultural relevance:** For 6 items, differences in the relevance of the content of the item to each culture were identified. In these cases, the translations were deemed appropriate, but the content of the item interacted with culture.

Sireci & Allalouf (2003)

Semoga Bermanfaat
